DOCUMENT RESUME

ED 439 685 IR 019 957

AUTHOR	Henrickson, Leslie
TITLE	Use of National Databases: Case Study on Census Data.
PUB DATE	1999-11-00
NOTE	11p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Case Studies; *Census Figures; *Comparative Analysis; *Data Analysis; Data Collection; Databases; Educational Research; Research Methodology
IDENTIFIERS	Census 1990

ABSTRACT

This paper serves as an example of how one may approach the use of a large database for comparative education purposes. Considerations include determining three aspects about the databases, and one case study. The three aspects include the type of database that is used, the method in which the data were collected and errors, and the structure of analysis. The paper draws on two authors in discussing the structure of analysis, Kish and Astin. Kish suggested two considerations used for analysis that help identify the strengths and weaknesses of different approaches: representation and realism. Astin used the I-E-O model, or Input-Environment-Output model, for design and analysis. A small case study using selected demographic and socioeconomic characteristics of various U.S. populations from the 1990 Census from the perspective of a comparative education researcher is described to illustrate how one's theoretical orientation toward a research project can guide analysis and interpretation of the data. (Contains a 15-item bibliography.) (AEF)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 439 685

LESLIE HENRICKSON
Social Sciences and Comparative Education
Graduate School of Education and Information Studies
University of California – Los Angeles
lhenrick@ucla.edu

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY HAS BEEN GRANTED BY

L. Henrickson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Use of National Databases: Case Study on Census Data

We live in an information rich society in which data on just about everything and everyone is available at the touch of a finger on the mouse. Information collected in myriad ways can be stored most easily for retrieval in the form of large databases. Vast stores of information, e.g. national databases on institutional structures, personal lifestyles and who talked to who, just as easily turn into gluts of information creating quagmires in the geography of knowledge. Unless one knows how to navigate through the "muck" of information the easiest course of action is avoidance. Such action would be unfortunate in this day and age when policy and funding decisions weigh heavily on the manipulation of statistical databases of these sorts.

The situation for education researchers is more pressing still. The ubiquity, advanced speed and power of personal computers makes it appear that access to potential knowledge can be available to anyone. For it used to be only highly trained scientists had access to and worked with databases. On the one hand, increased access to databases bodes well for those with good intentions who wish to right past wrongs. On the other hand, without proper training the potential to harness that information will always remain out of reach. Thus, avoidance is further structurally predetermined regardless of appearances.

Professional organizations call for change.[1] (Mechling, 93; Bottani, 1992). National organizations have begun to rectify this problem.[2] They promote the use of national databases through research funding

[1] A survey and discussion of leading-edge practitioners at the Conference on Advanced Computing in the Social Sciences showed two top priorities. "They felt that top priority for the future should be given to continued digitization of large databases and further improvements in user interfaces. These will make software and data more readily accessible to a much larger proportion of social scientists and the general population. Participants, in general, look forward to a future in which social science computing will rise with the general computing tide – that is, as computing grows exponentially more cost effective, the social

IR019957

BEST COPY AVAILABLE

and database training. This paper serves as an example of how one may approach the use of a large database for comparative education purposes. Considerations here include determining three aspects about the databases, and one case study. The three aspects are the type of database that is used, the method in which the data was collected and errors, the structure of analysis, and the example involves using 1990 census data from the perspective of a comparative education researcher. The example serves to illustrate how one's theoretical orientation toward a research project can guide analysis and interpretation of the data. Strengths and limitations to will be examined in light of the discussion on type, method and structure.

In general there are two types of databases, cross-sectional and longitudinal. Cross-sectional studies use data that are collected at one point in time over a particular unit of analysis about conditions, processes, outcomes, or other characteristics of interest. The unit of analysis for studies can include data from country studies, national, state, district and classroom investigations. Longitudinal studies use data that are collected over time in order to study stability or change of a unit of analysis about characteristics of interest. Although longitudinal studies provide the most comprehensive data, they are practically prohibitive due to high cost and difficulty in controlling for contingencies. Comparative research can be conducted on two or more single cross-sectional studies or one study that targets multiple population.

Data collection takes myriad forms that similarly can be collected cross-sectionally or longitudinally. The typical way of formulating and solving problems is usually related to the type of data collection. Qualitative and ethnographic data collection tends to formulate strategies for greater understanding of individual experiences. These strategies frequently involve face-to-face contact between researcher and subject, open-ended rather than closed-ended questions, unstructured rather than structured interview schedules, and samples are typically small. (Weber Cannon, et al. 115) However, input to a statistical database requires selective coding of the data that disallows the inclusion of some data and,

---

science community, along with society at large, will find increasingly pervasive and powerful applications for it."

[2] The National Center for Education Statistics and the National Science Foundation have developed rich data resources in the form of massive databases. However, the uses of these resources have, for a variety of reasons, lagged behind national-level data collection and organization. Institutions have lacked resources to develop the professional institutional capacity to access and to use the information. Lacking the use of detailed, quantitative national information, state and institutional-level post secondary leaders have not been able to readily place their planning problems and issues in a national data context. Toward this end their mission statements incorporate actions to foster the use of federal databases and to provide

therefore, is unavailable for analysis through these database techniques.[3] For example, lengthy narratives are excluded because databases generally have a limited field width and this type data is not amenable to statistical analysis. Quantitative data targets large populations and may be collected through surveys or through analysis of printed materials. It is presumed that the collected data fits a "variable oriented" method of analysis which treats theories as statements about the relative importance of variables as explanations of some result. The variable categories are generally pre-selected. (Ragin, et al, 1984) Cross-sectional surveys may be collected over whole populations, i.e. census, or may represent a sample of the population through random sampling techniques.

At least three types of error can occur: non-observation, measurement, and sampling. Non-observation errors include missing data because respondent declined to provide any or all information, or deliberately provided misleading information. Measurement errors could be due to a poorly constructed survey, variability in respondents or mistakes made by researcher in collection and analysis. Many sampling errors arise because the "part might not accurately represent the whole, faithfully reproducing its important characteristics." (Becker: 70) Admittedly, this is an oversimplification of a very complex problem, but usually random sampling is designed to deal with this difficulty. It is presumed that sampling errors have been eliminated in a census because the whole population has been included. However, census surveys include other sources of error like missing or incomplete data. An evaluative comment can be made on the relation of data collection and error. The attempt to conduct a census may result in a high proportion of missing or incomplete data, whereas a well-conducted survey with little missing data may yield data of higher quality, without bias, and with known sampling error.

I draw on two authors in discussing the structure of analysis, Kish and Astin. Kish suggested two considerations used for analysis, which help identify the strengths and weaknesses of different approaches, representation and realism. Representation refers to the generalizations drawn from analysis and how well

---

professional development for institutional researchers. See website: http://fundingopps2.cos.com/cgi-bin/getRec?id=14090

[3] There are non-statistical programming applications available to qualitative researchers with this sort of non-numerical, unstructured data that require ways of indexing, searching and theorizing. The most common being NUD*IST which is used for a wide variety of tasks from complex theory-constructing and testing to very quick analysis of small or large bodies of text in focus group summaries or open-ended answers in surveys. It is designed to automate much tedious work, swiftly "auto-coding" text and importing table data, and all processes can be done by command file. See website: http://qsr.latrobe.edu.au/

they apply to the population domain. Representation is best when sampling bias is controlled and randomization controls for sampling bias. Realism refers to the setting in which the investigation takes place. A natural setting may be preferable over a contrived laboratory setting because the environment may influence the outcomes. (Kish, 1987) Astin used the I-E-O model, or Input-Environment-Output, for design and analysis. The I-E-O model is both a conceptual and statistical tool used to measure the environment's effect on outcomes while controlling for input characteristics. Statistically, the I-E-O method entails temporally sequencing the variables under study as either input, or environment or outcome variables. Input variables are characterized as the initial point of time of the study. Inputs have a direct relationship with the environment and the outcomes. Therefore, input characteristics must be accounted for through controlling for them at the outset. Environment variables have to do with the individual's surroundings or experiences that take place between the measured times of the input and outcome. Input and environment are independent variables. The outcome measure is the dependent variable of the input and the environment. (Astin, 1993) The Kish and Astin models will be helpful in the case study.

The question can be raised how does one do comparative education research using large databases? A small case study using selected demographic and socioeconomic characteristics of various U.S. populations from the 1990 Census will illustrate how a researcher might begin a comparative analysis. A census is a cross-sectional database over a whole population and, therefore, is subject to the limitations and errors mentioned above. Using the information discussed above as well as some information on two theoretical orientations will help guide the discussion. Table 1 shows selected characteristics on five populations. I make a few assumptions about this data that affect my analysis. First, the percentage total adds to 96.6% of the total population and I assume the missing 3.4% refers to the Latino population. Second, Asian Pac/Is is ambiguous and is not disaggregated as the Latino/Non-Latino distinction. Therefore, I assume that Asian Pac/Is includes at least all of the following: Asian/PI, Chinese, Filipino, Japanese, Korean, Asian Indian and Vietnamese.

The first thing to consider for any study is the research question. A comparative study requires at least two populations of which one is the base population. Which populations are chosen depends on the research question. The purpose of this study is to illustrate how the census has been used and to show how it can be used in different ways. Three populations were selected: Chicana/o, Asian Pac/Is and White.

Although this is a national database, the populations I am ultimately interested in are California school children and the three populations represent the three largest student populations in California schools.[4] The White population will be used as the base because it represents 75.8% of the total population.

There are five variables of demographic information: total number of each population, percent of total population, median age of the population and the mean family size, the degree of English fluency. There are two variables of educational information, the percent having completed a high school degree and the percent having completed a bachelors degree. There are five variables of socioeconomic information: the percent female or male who are professional managers, the percent female or male unemployed, the median family income.

Table 1. Selected Demographic and Socioeconomic Characteristics of Various U.S. Populations: 1990

	Non-Latino				
	Chicana/o	Native American	**Asian Pac/Is**	Black	**White**
Number-Millions (Total US = 248.7)	**13.4**	1.9	**7.0**	29.3	**188.4**
% of Population	**5.4**	0.8	**2.8**	11.8	**75.8**
Median Age	**23.8**	27.3	**30.3**	28.4	**34.9**
Mean Size of Family	**4.08**	3.56	**3.74**	3.46	**3.03**
Does not Speak English "Very Well" (5yrs+)	**38.9**	8.4	**38.9**	1.7	**1.8**
%4 years High School or more (25yrs+)	**44.2**	66.0	**77.7**	63.3	**79.1**
%BA Degree or more (25yrs+)	**6.3**	9.4	**37.0**	11.4	**22.0**
%Manager/professional occupation: Female	**15.2**	22.1	**28.4**	21.4	**29.8**
%Manager/professional occupation: Male	**9.4**	15.7	**33.2**	14.6	**27.5**
Unemployed: Male	**11.7**	13.1	**5.4**	12.1	**4.8**
Unemployed: Female	**10**	15.6	**5.0**	13.7	**5.1**
Median Family Income	**$24,119**	$21,661	**$41,446**	$22,466	**$37,628**

Following Astin's I-E-O model we can organize the variables into a sequential flow. Grant that the breakdown will be subjective for illustrative purposes, and bar arguments that the relationships between

[4] NCES-CA Dept. of Ed. 1994: Hispanic (Chicana/o and Latino) 37.9%, Asian Pac/Is 11.2%, White 41.4%

variables are cyclical. The model requires input variables because they are related to environments and outcomes. It is important to control for inputs because any observed relationship between environments and outcomes might well be reflecting the effects of inputs rather than the true effect of an environment. (Astin: 265) Input variables would be median age, mean size of family, English fluency, and median family income. Along a time continuum education activities happen next. The environment variables would be percent completed high school degrees and percent completed bachelors degrees. The outcome variables would generally be employment-related items. The output variables would be the percent female/male as professional managers and the percent female/male unemployed.

Statistical analysis comprised of step-wise regression analysis could be run on the variables' raw data in the sequential I-E-O order to find correlations between variables. Step-wise regression analysis is a mathematically legitimate type of statistical procedure to carry out on large databases like the census. When the number of cases is small the regression results are not reliable. Whether the representation or realism of the variables appropriately captures the characteristics of interest is highly dependent on the research question and the particular study. A regression analysis yields correlation between variables. Astin emphasizes that "while it is true that you can't prove causation with correlational data, you most certainly can make causal inferences from such data." (Astin: 29) The emphatic point behind this statement is that traditionally and historically large databases have been used to "prove facts" so as to shore up some preconceived world view.[5] Often these "facts" reinforced discriminatory practices and false stereotypes as I discuss below.

Astin's point that I want to stress is that large databases are useful to point in fruitful directions toward further research. However, one cannot stop at the first sign of a correlation and claim causal proof.

---

[5] This comment is taken from Astin's 1991 text and seems at odds with Don Nakanishi's observation that Astin himself had made an unfounded assumption and assertion about Asian American students ten years earlier. In this case, it was the failure to include data in the database that shored up false assumptions. Nakanishi's states, "In two of the more highly regarded comparative studies of minority students – Alexander Astin's *Minorities in Higher Education* (1982) and John Ogbu's *Minority Education and Caste* (1978) – Asian Americans were not included in the data collection and analysis because they were not considered to be 'educationally disadvantaged' like other non-white minority groups. Indeed, there is much to suggest that the admissions debate might not have become so explosive if there had been a body of empirical knowledge that all parties to the dispute could have used to test or verify their largely unfounded assumptions and assertions about Asian American students." The inconsistency between Astin's previous omission and his assessment on correlational data does not detract from the usefulness of his method. More to the point, it warns us to always be on guard.

This point can be illustrated as follows. Compare the three populations to one another. We find that the Chicana/o and the Asian Pac/Is data are similar for the demographic variables of percent of the population, mean size of family and English fluency. That is, compared to the base population both represent a small percent of the population, have larger families and have about 40% of the populations do not speak English well. We find that the Asian Pac/Is and the White data are similar on the educational and socioeconomic variables of percent completed high school, percent female/male professional manager, percent female/male unemployed. That is, both have about 80% high school completion, 29% of female/male in professional positions, both have low unemployment rates and both earn about $40,000/year. On all these accounts the Chicana/o population is roughly half.

Using this first cut comparison, one could say that controlling for input variables, since Chicana/o and Asian Pac/Is were similar, that the outcomes of high employment and earnings were due to the environment of educational attainment. If one were looking for proof that a minority could pull themselves up by their bootstraps through rugged individualism or through an Asian culture which values education and achieve the American dream, then the Asian Pac/Is appear to the be the "model minority". Historically, researchers focused on proving a connection between the role of education and occupational achievement were oriented toward theories of adaptation, either assimilation theory or human capital theory. (Barringer, et al. 148) Assimilation theory proposes that the longer a minority was in a region the closer their income level will be to the base population. Human capital theory proposes that the greater educational attainment of a minority will increase their income level.

Such proof that success could be attained shifted the empirical lens toward the "deficit" in occupational success in the Chicana/o population. Researchers looked for some rationale that explained the outcome disparity given the similar initial starting conditions as Asian Pac/Is. Their focus pointed toward the environmental variable of education, in this case, the inadequate educational achievement and attainment of Chicana/os. There has been a long history of various conceptual frameworks advanced to explain this school failure, which coalesce into a model called the deficit thinking paradigm. Deficit thinking "posits that students who fail in school do so because of alleged internal deficiencies (such as cognitive and/or motivational limitations) or shortcomings socially linked to the youngster – such as

familial deficits and dysfunctions. Given the endogenous nature of deficit thinking, systemic factors are held blameless in explaining why some students fail in school." (Valencia, 1997)

Instead of using correlations between variables to shore up faulty reasoning and racist theories, such correlations should point to more fruitful lines of research to disaggregate the complexities hidden within large databases. For one thing, the use of databases to support these theories fail to account for historical conditions that make the correlations reasonable, e.g. structural limitations like school segregation and inequities in school finance. Moreover, a statistical analysis that compares to a baseline with predetermined categories presumes the baseline categories legitimately hold for each population. On the contrary, from a minority populations' perspective important categories specific to that population will be omitted or misrepresented. Thus, there is no realism in the analysis. For example, I have assumed that the Asian Pac/Is refers to a multitude of different ethnic minorities that may have completely different background. It is misleading and unfair to presume those Khmer refugees new to this country would be statistically significant to well-educated Japanese immigrants.

Valencia and Solorzano show that the historical presence of deficit thinking has contemporary overtones. They propose to push the research envelope. "In our analysis of these contemporary deficit thinking expressions, we need, at times, to go beyond educational thought and practice in that such perspectives have worked their way into the broader context of American social though regarding race, ethnic, gender and class relations." (Valencia and Solorzano, 1997) One way to begin to understand individual differences along the lines of gender, race and class would be through a more nuanced methodology called feminist standpoint theory. Standpoint takes the analysis of gender and social power as its starting place. Proponents of this position believe that situatedness within a particular gender or disadvantaged social group produces an epistemological advantage in interpreting truth, e.g. in terms of sociopolitical issues. The objectivity of the oppressed is privileged in a process of constructing an emancipatory reformation. (Harding, 1991) Feminists have enlarged the critique of "scientific objectivity" to argue that it is not just the content of traditional disciplines that is problematic, but the methods that have been used to gather knowledge about women and minorities as well. Toward this end, qualitative and ethnographic methods have provided valuable insight into women's lives by putting their

lived experiences first and foremost, thereby, decentering the white, Western, patriarchal pre-determined categories of what had constituted legitimate knowledge.

Some avenues of inquiry to explore with the given census data set include the following. Break down the monolithic Asian Pac/Is category to show the variegated nature of the sub-populations. Look into the relationship of English fluency among sub-populations, education and employment. Some of these Asian Pac/Is sub-populations will be more similar to Chicana/o experience than others. Explore through standpoint methodology these sub-population experiences looking at power relations, structural limitations and socio-cultural influences. It is interesting to note that both the Chicana/os and Whites have a higher percentage of women in professional management positions than Asian Pac/is, holding unemployment constant across populations. A study on the similarities between Chicana/o and White women managers raises interesting questions about gender and education. To explore gender and education calls for disaggregating the educational statistics along gender lines and exploring through both qualitative and quantitative means the nature of this more complex relationship.

We back off from the expectation that large databases will provide proof and rather look toward the correlations between variables as road signs of fruitful lines of further inquiry. Certainly, there are limitations to these alternative forms of inquiry. The worst limitation of all is to rely univocally upon one method as the true means of determination. Large databases allow us to see certain aggregated trends and relationships between gross variables in the form of predetermined categories. For example, they can have construct validity. Language census scores within a bilingual population correlate more substantially with wholistic measures because they are self-reported measures on usage and proficiency. (Fishman: 648) Yet, it has been reported that for the 1990 census the non-response rate increased by 15% which introduces error in the results. (Singer, et al: 467) Translating the resulting numbers into meaningful actions by real people is not always an easy exercise. Further, the questions these analyses answer are not always the ones that people want to know. Interviews and ethnographic studies embody a different way of thinking about the world. The sample size is small, not easily generalizable but it creates a closer link between researcher and researched, acknowledges the role of values in research and on the whole is more contextual and subjective. No method alone can do the trick, it is through multiple perspectives that we will gain greater understanding. (Kellner, 1995: 98-101)

Bibliography

1. Astin, Alexander W. Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation in Higher Education. Phoenix, Arizona: Oryx Press; 1993.

2. Barringer, Herbert R.; Takeuchi, David T.; and Xenos, Peter. Education, Occupational Presitge, and Income of Asian Americans. The Asian American Educational Experience: A Source Book for Teachers and Students. New York: Routledge; 1995;(Nakanishi, Don T. and Yamano Nishida, Tina.

3. Becker, Howard. Tricks of the Trade: How to Think About Your Research While You're Doing It. Chicago: University of Chicago Press; 1998.

4. Bottanni, N.; Duchene, C.; and Tuijnman, A. Education at a Glance: OECD Indicators. Paris: OECD; 1992.

5. Fishman, Joshua A. and Terry, Charles. The Validity of Census Data on Bilingualism in a Puerto Rican Neighborhood. American Sociological Review. 1969; 34(5) 636-650.

6. Harding, Sandra. Whose Science? Whose Knowledge? Thinking from Women's Lives. Ithaca, New York: Cornell University Press; 1991.

7. Kellner, Douglas. 1995. *Media Culture. Cultural Studies, Identity and Politics Between the Modern and the Postmodern*. London: Routledge.

8. Mechling, Jerry. Priorities for Social Science Computing: A Survey of Leading Edge Practitioners. Social Science Computing Review. 1991 Spring; 9(1) 93-103.

9. Moore, David S. and McCabe, George P. Introduction to the Practice of Statistics. 2nd ed. New York: W. H. Freeman and Company; 1993.

10. Nakanishi, Don T. A Quota on Excellence? The Asian American Admissions Debate. The Asian American Educational Experience: A Source Book for Teachers and Students. New York: Routledge; 1995; pp. 273-284. (Nakanishi, Don T. and Yamano Nishida, Tina.

11. Ragin, Charles C.; Meyer, Susan, and Drass, Kris. Assessing Discrimination: A Boolean Approach. American Sociological Review. 1984; 49.221-234.

12. Singer, Eleanor; Mathiowetz, Nancy A.; and Couper, Mick P. The Impact of Privacy and Confidentiality Concerns on Survey Participation: The Case of the 1990 U.S. Census. Public Opinion Quarterly. 1993 Winter; 57(4) 465-482.

13. Valencia, Richard R. Conceptualzing the notion of Deficit Thinking. The Evolution of Deficit Thinking: Educational Thought and Practice. London: Falmer Press; 1997; pp. 1-12. (Valencia, Richard R.

14. Valencia, Richard R. and Solorzano, Daniel G. Contemporary Deficit Thinking. The Evolution of Deficit Thinking: Educational Thought and Practice. London: Falmer; 1997; pp. 160-210.

15. Weber Cannon, Lynn; Higginbotham, Elizabeth, and Leung, Marianne L. A. Race and Class Bias in Qualitative Research on Women. Beyond Methodology: Femiist Scholarship as Lived Research. Bloomington: Indiana Press University; 1991; pp. 107-118.